Trường Đại học Công nghệ Thông tin (UIT)
Đại Học Quốc Gia Tp. HCM

BÀI TẬP 1

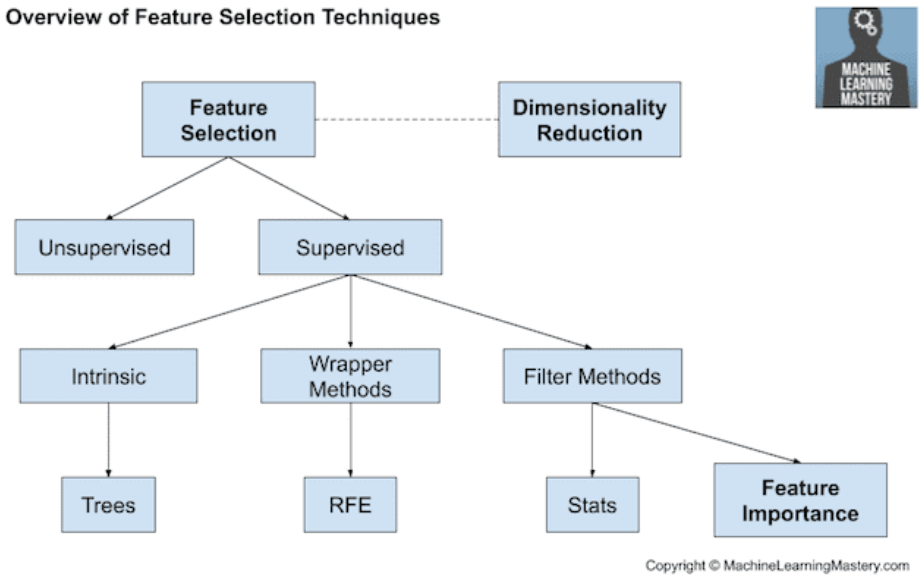
Môn học: NT522 – Phương pháp học máy trong An toàn thông tin

Tên chủ đề: **Feature Engineering in Malware Detection**

Mã môn học: NT522

1. NỘI DUNG THỰC HIỆN

Feature Engineering (Xử lý đặc trưng) *là một giai đoạn không thể thiếu trong quá trình phát triển bất kỳ một hệ thống thông minh nào. Mặc dù hiện nay chúng ta có rất nhiều các phương pháp mới như học sâu, siêu mô hình hỗ trợ học máy tự động (automated machine learning), tuy nhiên với mỗi vấn đề cụ thể cần giải quyết luôn có những đặc trưng quan trọng hơn, có giá trị hơn để quyết định hiệu suất hệ thống. Feature Engineering vừa là nghệ thuật cũng là một môn khoa học và đây là lý do để các nhà khoa học dữ liệu thường dành tới 70% thời gian của họ cho giai đoạn chuẩn bị dữ liệu trước khi xây* dựng *mô hình. Đây là quá trình biến đổi dữ liệu thành các đặc trưng đóng vai trò là đầu vào cho các mô hình học máy. Các đặc trưng được xử lý tốt sẽ nâng cao hiệu suất của mô hình. Các đặc trưng cũng ảnh hưởng rất lớn bởi vấn đề cần giải quyết trong bài toán tương ứng.*



Hình 1. Các kỹ thuật chọn lựa thuộc tính cho quá trình huấn luyện mô hình học máy

Trong lĩnh vực an toàn, bảo mật thông tin, nhiều nghiên cứu đã được thực hiện để phát hiện, phân loại phần mềm độc hại dựa trên việc học máy sử dụng các đặc trưng từ chương trình phần mềm được trích xuất bằng phương pháp phân tích tĩnh. Trong bài tập này, hãy thử xem xét bài toán phân biệt phần mềm độc hại và chương trình lành tính bằng cách tìm hiểu các

đặc điểm đặc trưng của chúng, chẳng hạn như thông tin tổng quan (general information) hay các lời gọi hàm được truy nhập trước (imported functions). Để hệ thống này có khả năng áp dụng trong thực tế, cần có sự cân bằng tốt giữa độ chính xác, thời gian học và kích thước dữ liệu của các thuộc tính sử dụng. Mặc dù phương pháp chỉ sử dụng một tập hợp con các tính năng có thể giảm thời gian và kích thước dữ liệu cần thiết cho việc xây dựng mô hình hiệu quả, nhưng việc chọn một tập hợp con thích hợp từ tổng số các đặc trưng không phải là điều đơn giản. Hãy tìm hiểu, và thực nghiệm về kỹ thuật xử lý dữ liệu, lựa chọn tính năng trong phát hiện phần mềm độc hại dựa trên phương pháp học máy giám sát. Trong ngữ cảnh bài tập này, tập dữ liệu Ember được chọn làm dữ liệu mục tiêu để xác định các tổ hợp đặc trưng phù hợp tương ứng với độ chính xác, thời gian học và kích thước dữ liệu cho việc xây dựng mô hình nhận diện mã độc.

Yêu cầu: Áp dụng các kỹ thuật rút trích, chọn lựa, xử lý đặc trưng (feature engineering) để chọn ra bộ thuộc tính phù hợp, tối ưu để sử dụng cho bài toán nhận diện, phân loại mã độc trên **tập dữ liệu Ember (phân loại nhị phân -binary classification)**.

a) Phân tích chung về bộ dữ liệu (tổng số nhãn, tổng số thuộc tính, phân phối của các nhãn dữ liệu…). Yêu cầu lập trình code để in các thông tin này ra màn hình quan sát.
b) Mỗi nhóm tìm hiểu và thực hiện đầy đủ **03** nhóm phương pháp: **Wrapper, Filter, Intrinsic** trong việc chọn lựa thuộc tính như các kỹ thuật sau:
– Sử dụng các chiến lược đánh giá tầm quan trọng của thuộc tính (Feature Importance)
– Sử dụng chiến lược thống kê mối tương quan giữa các thuộc tính (Correlation Statistics): Pearson's Correlation Coefficient, Chi-Squared, …
– Sử dụng phương pháp dựa trên cấu trúc cây (Tree)
– Sử dụng chiến lược Recursive Feature Elimination (RFE) trong chọn lựa thuộc tính
c) Thực hiện thí nghiệm sử dụng bộ thuộc tính đã chọn lựa ***để huấn luyện 02 mô hình học máy (01 mô hình ML và 01 mô hình DL cơ bản)***, đánh giá tầm ảnh hưởng của phương pháp rút trích và chọn lựa đặc trưng đối với hiệu năng của mô hình.
VD: Nhóm sinh viên có thể so sánh trade-off giữa thời gian training và độ chính xác (Accuracy, Pre, Re, F1) của 02 mô hình trên, trong điều kiện thực hiện và không thực hiện thao tác kỹ thuật xử lý đặc trưng (feature engineering).

Lưu ý:

+ Mô tả rõ ràng cách tiền xử lý dữ liệu và thành phần dữ liệu nào được chọn để làm thí nghiệm.
+ Các nhóm có thể tham khảo tài liệu: "Y. Oyama, T. Miyashita and H. Kokubo, *"Identifying Useful Features for Malware Detection in the Ember Dataset*," 2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW), 2019, pp. 360-366, doi: 10.1109/CANDARW.2019.00069." để thực hiện.

Tài nguyên:

- Bộ dữ liệu mã độc: **Ember dataset** (tùy chọn sử dụng toàn bộ hoặc chọn một phần trong tổng số các tập *train1/train2/train3/train4/train5* cho quá trình huấn luyện và tập *test* cho quá trình kiểm tra).

- Trong điều kiện hạn chế về tài nguyên, các nhóm có thể sử dụng một phần của bộ dữ liệu trên (ví dụ: 40%, 50%, 60%, 70%, … tập tương ứng). Khuyến khích sử dụng nhiều dữ liệu hơn trong điều kiện cho phép.
- Liên kết tải: https://ember.elastic.co/ember_dataset_2018_2.tar.bz2

Hình 2. Tập dữ liệu Ember 2018

2. GỢI Ý – THAM KHẢO

Một số gợi ý thực hiện:

- EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models: https://arxiv.org/pdf/1804.04637.pdf
- Identifying Useful Features for Malware Detection in the Ember Dataset: https://ieeexplore.ieee.org/document/8951564
- Dataset:

  https://paperswithcode.com/dataset/ember

  https://github.com/elastic/ember
- Bài giảng và sách liên quan tới chủ đề "Feature Engineering" của môn học NT522.
- Blog Data Preparation for Machine Learning (7-Day Mini-Course): https://machinelearningmastery.com/data-preparation-for-machine-learning-7-day-mini-course/
- Blog "How to Calculate Feature Importance With Python": https://machinelearningmastery.com/calculate-feature-importance-with-python/
- Thư viện hỗ trợ scikit-learn, …

Sinh viên đọc kỹ yêu cầu trình bày bên dưới trang này

YÊU CẦU CHUNG

- Sinh viên tìm hiểu và thực hiện bài tập theo yêu cầu, hướng dẫn.
- Nộp báo cáo kết quả chi tiết những việc (**Report**) bạn đã thực hiện, quan sát thấy và kèm ảnh chụp màn hình kết quả (nếu có); giải thích cho quan sát (nếu có).
- Sinh viên báo cáo kết quả thực hiện và nộp bài.
- **Nộp kèm Link chứa code Notebook trên Google Colab, đã chạy và hiển thị kết quả sau khi huấn luyện mô hình.**

Báo cáo:

- File .PDF. Tập trung vào nội dung, không mô tả lý thuyết.
- Nội dung trình bày bằng Font chữ Times New Romans/ hoặc font chữ của mẫu báo cáo này (UTM Neo Sans Intel/UTM Viet Sach)– cỡ chữ 13. Canh đều (Justify) cho văn bản. Canh giữa (Center) cho ảnh chụp.
- Đặt tên theo định dạng: [Mã lớp]-ExeX_GroupY. (trong đó X là Thứ tự Bài tập, Y là mã số thứ tự nhóm trong danh sách mà GV phụ trách công bố).

  *Ví dụ: [*NT101.K11.ANTT*]-Exe01_Group03.*
- Nếu báo cáo có nhiều file, nén tất cả file vào file .ZIP với cùng tên file báo cáo.
- Không đặt tên đúng định dạng – yêu cầu, sẽ **KHÔNG** chấm điểm bài nộp.
- Nộp file báo cáo trên theo thời gian đã thống nhất tại courses.uit.edu.vn.

Đánh giá:

- Hoàn thành tốt yêu cầu được giao.
- Có nội dung mở rộng, ứng dụng.

Bài sao chép, trễ, … sẽ được xử lý tùy mức độ vi phạm.

HẾT